The multiscale coarse-graining method. VI. Implementation of three-body coarse-grained potentials

Luca Larini,[1,a)] Lanyuan Lu,[1,2,a)] and Gregory A. Voth[1,2,b)]
[1]*Department of Chemistry and Center for Biophysical Modeling and Simulation, University of Utah, 315 S 1400 E, Salt Lake City, Utah 84112, USA*
[2]*Department of Chemistry, James Franck Institute, and Computation Institute, University of Chicago, 5735 S. Ellis Ave., Chicago, Illinois 60637, USA*

(Received 17 December 2009; accepted 25 March 2010; published online 26 April 2010)

Many methodologies have been proposed to build reliable and computationally fast coarse-grained potentials. Typically, these force fields rely on the assumption that the relevant properties of the system under examination can be reproduced using a pairwise decomposition of the effective coarse-grained forces. In this work it is shown that an extension of the multiscale coarse-graining technique can be employed to parameterize a certain class of two-body and three-body force fields from atomistic configurations. The use of explicit three-body potentials greatly improves the results over the more commonly used two-body approximation. The method proposed here is applied to develop accurate one-site coarse-grained water models. © *2010 American Institute of Physics*. [doi:10.1063/1.3394863]

I. INTRODUCTION

Coarse-grained (CG) models are becoming essential instruments for both theoretical and computational molecular science (see, e.g., Ref. 1).[1,2] Theoretical models in general have contributed many new insights, for example, in polymer[3–5] and liquid crystal physics.[6] Computationally inexpensive CG models can similarly describe phenomena such as protein structure,[7] signal transduction,[8] and virus capsids,[9] which otherwise would be beyond the reach of standard molecular dynamics (MD) techniques,[10–12] which employ conventional all-atom force fields. Many methodologies have been developed to build CG force fields such as the Boltzmann inversion,[13,14] iterative Boltzmann inversion,[15] reverse (or inverse) Monte Carlo (RMC),[16,17] and multiscale coarse-graining (MS-CG)[18–22] methods. Until now, these models have been used to construct pairwise additive potentials, without considering explicit multibody interactions between the CG sites. Methods such as Boltzmann inversion can, in principle, be used for constructing multibody interactions starting from multibody distribution functions. However, such methods assume that each distribution can be fitted independently from the others. Generally, this assumption cannot be considered valid for every system. RMC in principle can take into account correlations among different distributions.[16] However, due to the prohibitive computational cost that comes with introducing three-body potentials into RMC,[16] no such RMC application has been reported. In the present paper, an extension of the MS-CG methodology will be proposed, which can effectively and efficiently develop three-body CG potentials. The resulting models are compared to models obtained using the standard two-body MS-CG method without three-body potentials, in order to clearly identify the benefit of such potentials. To the authors' best knowledge, this work represents the first time that a three-body CG potential has been systematically derived from the underlying atomistic interactions.

The MS-CG methodology was initially proposed by Izvekov and Voth.[18,19] The idea behind the MS-CG method is that the effective CG forces between different sites originate from a statistical averaging of their atomistic interactions. For this reason, it is expected that certain properties of a successful CG model can compare nearly exactly with the underlying atomistic system. Obviously, as a change of resolution is involved in the coarse-graining process, some sort of average over the irrelevant degrees of freedom is required. In the original formulation of MS-CG, only pairwise additive potentials were employed and cubic splines were used to accurately fit the effective CG forces obtained from averaging over atomistic resolution MD data. A complete theoretical foundation for the MS-CG technique was then reported by Noid *et al.*[20–22] In the latter work, it is shown that the degrees of freedom that are not to be included in a CG model can be properly averaged over via a variational algorithm. In particular, it was shown how a simple pairwise additive assumption for the CG effective force field is able to account for some higher order correlations that are usually present in the original atomistic system.[20] However, due to numerical difficulties, the explicit development of three-body effective CG potentials, which might lead to a more accurate description of multisite correlations, has to date not been attempted.

The idea of matching force at the atomistic level was first proposed by Ercolessi and Adams[23] in the form of a force-matching (FM) methodology. In that case, the force acting on each atom was derived from *ab initio* quantum calculations that were used to parameterize a classical force field. This FM methodology was later reformulated and made applicable to liquid and soft condensed matter by the

[a)]These authors contributed equally to this work.
[b)]Author to whom correspondence should be addressed. Electronic mail: gavoth@uchicago.edu.

use of a linear basis set, which made the problem numerically tractable for such problems.[24,25] By contrast, the MS-CG approach was first developed[18,19] to systematically reduce the number of degrees of freedom within a given system (i.e., to coarse grain it), whereas no change in the number of degrees of freedom is involved in the methodology employed by Ercolessi and Adams. Also (and importantly) in Refs. 20 and 21, a theoretical statistical mechanical basis has been given for the MS-CG method that clearly demonstrates its differences from the original FM procedure by Ercolessi and Adams.[3] Thus, the MS-CG method is significantly different the FM approach given by Ercolessi and Adams,[23] despite persistent confusion in some of the coarse-graining literature that continues to refer to the MS-CG approach as FM. MS-CG should, in fact, be more accurately considered to be a kind of statistical mechanical "force renormalization."

In context of effective condensed phase interactions, the importance of three-body intermolecular potentials for describing the properties of materials has been extensively reported in the condensed matter physics literature. Their effects appear even at an atomistic level, where three-body interactions are essential, e.g., for representing the structure of covalently bonded atoms,[26–32] as well as for accurately reproducing the equation of state of noble gases such as ^{4}He (Ref. 33) and argon.[34] At larger scales, three-body interactions appear to play an important role in defining the packing properties of nanocrystals[35] and proteins.[36,37] Even dynamical properties such as the rate of folding in proteins may be affected by the presence of three-body interactions.[38] The use of three-body potentials to correctly model bonded interactions is widely established. On the other hand, the use of three-body potentials to describe nonbonded interaction is usually avoided, mainly for two reasons. The first reason is due to the fact that they can be computationally expensive. Another reason is that they require the parameterization of multiple parameters. However, as will be shown in the present work, three-body potentials are suitable for accurately modeling effective short range nonbonded interactions in CG models. In this case, the fact that they are more computationally expensive than the two-body terms is compensated for by the short range of the three-body effective interaction and by the reduction of degrees of freedom in the CG system, making their implementation competitive with traditional all atom force fields.

In the present work, a new algorithm is therefore proposed to systematically build MS-CG force fields having explicit three-body CG potentials. In turn, the new method is used to study the relevant properties of the Single Point Charge/Extended (SPC/E) water[39] model. The relevance of three-body interactions for reproducing water properties at the CG level has been convincingly demonstrated by Molinero and Moore.[31] In their work, the functional form used for the CG water force field was chosen to be the Stillinger–Weber (SW) potential,[32] originally developed for silicon. However, over the years, this potential has shown its flexibility in successfully reproducing the properties of other systems.[26–32] Inspired by this body of work, a combination of this model with the MS-CG approach will be developed here in order to improve upon the MS-CG force field.

The remainder of this paper is structured as follows. Section II is devoted to a detailed description of the new algorithm. In Sec. III the new method is employed to build MS-CG model for SPC/E water.[39] A comparison with a simpler pairwise MS-CG approximation is also provided in the same section. Conclusions are then given in Sec. IV.

II. METHODS

A. Coarse-graining procedure

The basic idea of the MS-CG method is as follows.[18–22] The original atomistic system is first mapped to a reduced system having a less complicated structure. Sets of atoms on the same molecule are grouped together to define coarse grained sites. A set of basis functions for the CG force function is constructed, and the CG force function is represented as a linear combination of the basis functions. The coefficients of the basis functions are obtained from a variational principle that uses atomistic simulation data. A detailed description of both the theoretical background[20–22] and the actual numerical implementation[40] of the method can be found elsewhere. The previous implementations of the method[18–20] for constructing nonbonded CG interactions use basis functions that describe only pairwise additive forces between the CG sites. In the present work, an algorithm is proposed to extend the MS-CG methodology to develop three-body CG force fields.

B. Limitations of the two-body approximation in CG modeling

The coarse graining of dense systems of interacting particles involves dealing with a complicated multibody potential of mean force. From a practical point of view, directly computing and utilizing such multibody force fields is not possible, so approximations must be introduced. Usually such approximations involve the modeling of pairwise interactions of such a complicated mean force field, under the implicit assumption that higher order components, such as three-body terms, can be neglected. However, when dealing with complex molecular structures, this approximation may not be valid. Properties such as hydrophobicity are intrinsically due to multibody interactions,[41] requiring a more detailed approach to the problem which involves going beyond traditional approximations.

Another very common approximation that is used when modeling two-body potentials is the spherical approximation, where the interaction between two centers depends only on their distance, but not on their mutual orientation. Although, in principle, there is no need for such an approximation, it is of great practical utility. However, this approximation cannot be suitable for every system. In CG modeling, where part of complex molecules are usually coarse grained into a single interaction site, the loss of information regarding their mutual angular orientation can lead to problematic behavior. An example of this can be found in liquid crystals,[6] where their most basic properties cannot be reproduced without taking into account their angular anisotropy.

As explained above, building "rigorous" (as opposed to *ad hoc*) CG models requires the construction of an approximate multibody potential of mean force for the CG variables. The underlying concept is that only the most essential properties of the system under examination will be retained in the CG representation. For these reasons, a set of quantities that are considered most relevant for the system of interest should be selected and well reproduced in the corresponding CG representation. A quantity that is commonly chosen is the radial distribution function (RDF) of key CG variables in the system. This function has proven to be of central importance in the theory of liquids,[42] and methods such as RMC specifically target this quantity.[16,17] The RMC approximation commonly assumes that two-body CG effective potentials are sufficient to construct CG models that correctly reproduce the RDF.[43] On the other hand, for the MS-CG method the target quantity is the multidimensional potential of mean force itself for the CG variables. In the original MS-CG methodology, the true multidimensional potential of mean force is approximated by only two-body nonbonded contributions to represent the effective CG force field. Indeed, effective CG force fields generated using different methodologies will lead to different results for the RDF, as the approximations involved can be very different. For example, RDFs from the MS-CG method are not likely to be as accurately reproduced as they are in RMC since the latter method is designed to reproduce the RDF.[43] As a result, the RDFs from the MS-CG approach are a suitable choice for assessing the quality of an approximate MS-CG force field, while the RMC method will produce an accurate RDF but it will not likely reproduce other quantities well such as three-body correlations in the CG variables, and therefore RMC requires an additional strategy for validation beyond a simple comparison to the correct RDF. For the reasons described above, a generalization of the MS-CG approach to include three-body potentials is proposed here.

C. Generalization to three-body CG potentials

Three-body potentials depend on the coordinates of three interacting sites and it can be written in a general form as $U(\theta_{jik}, r_{ij}, r_{ik})$, where i is the index for the central site and j and k are the other two sites. For the purpose of this paper, we assume that this potential can be expanded as

$$U(\theta_{jik}, r_{ij}, r_{ik}) = \sum_{\nu} c^{(\nu)} A^{(\nu)}(\theta_{jik}) S^{(\nu)}(r_{ij}) S^{(\nu)}(r_{ik}), \tag{1}$$

where, for the νth basis function, $c^{(\nu)}$ is its coefficient, $A^{(\nu)}(\theta_{jik})$ denotes an angular dependent term, and $S^{(\nu)}(r_{ij})$ and $S^{(\nu)}(r_{ik})$ are two distance dependent terms.

In the following, some approximations to the original expansion in Eq. (1) are employed in order to make the method of practical applicability. The original function will be approximation by only one term of the expansion, such that

$$U(\theta_{jik}, r_{ij}, r_{ik}) \approx A(\theta_{jik}) S(r_{ij}) S(r_{ik}). \tag{2}$$

In order to obtain the force, the derivative of the potential is calculated as[44]

$$\begin{aligned} -\frac{\partial}{\partial r_l^{\alpha}} U(\theta_{jik}, r_{ij}, r_{ik}) &= -S(r_{ij}) S(r_{ik}) \frac{\partial}{\partial r_l^{\alpha}} A(\theta_{jik}) - A(\theta_{jik}) S(r_{ik}) (\delta_{lj} - \delta_{li}) \\ &\times \frac{r_{ij}^{\alpha}}{r_{ij}} \frac{\partial}{\partial r_{ij}} S(r_{ij}) - A(\theta_{jik}) S(r_{ij}) (\delta_{lk} - \delta_{li}) \\ &\times \frac{r_{ik}^{\alpha}}{r_{ik}} \frac{\partial}{\partial r_{ik}} S(r_{ik}), \end{aligned} \tag{3}$$

where $l \in \{i, j, k\}$, α indicates one of the x, y, and z components, and δ is the Kronecker delta function. The goal of the three-body MS-CG algorithm is to construct a set of basis functions for the three-body potential and minimize the residual in the MS-CG variational principle.

Another approximation introduced is that the three-body potential will act only on a very short range, roughly the first solvation shell. As computing nonbonded three-body potentials is extremely expensive, this assumption is of practical relevance.

In the rest of the paper, the three-body potential will be used as a correction to the spherical approximation introduced by the use of two-body potentials alone. Thus, it is assumed that the major correction to a simple pairwise approximation will come to an accurate fitting of the angular term $A(\theta_{jik})$. The functions $S(r_{ij})$ and $S(r_{ik})$ will be used to guarantee that its effect wears off gently after the cutoff.

Furthermore, from a more technical point of view, in the original MS-CG formulation involving only pairwise potentials, all the functional forms involved were fitted with spline functions. However, if spline functions are used for both angular and distance dependent terms in the three-body term, the MS-CG least-squares problem becomes a *nonlinear* equation. Therefore, to take advantage of linear least square methods, approximations must be implemented, such as specifying a portion of the three-body potential before the application of the MS-CG procedure.

For these reasons, for the three-body part of the potential the widely used SW (Ref. 32) potential is chosen with the form

$$\begin{aligned} U(\theta_{jik}, r_{ij}, r_{ik}) = \lambda_{jik} \varepsilon_{jik} (\cos\theta_{jik} - \cos\theta_0)^2 \\ \times \exp\left(\frac{\gamma_{ij}\sigma_{ij}}{r_{ij} - a_{ij}\sigma_{ij}}\right) \exp\left(\frac{\gamma_{ik}\sigma_{ik}}{r_{ik} - a_{ik}\sigma_{ik}}\right), \end{aligned} \tag{4}$$

For the specific case of one component system such as water, it can be simplified as

$$\begin{aligned} U(\theta_{jik}, r_{ij}, r_{ik}) = \lambda\varepsilon (\cos\theta_{jik} \\ - \cos\theta_0)^2 \exp\left(\frac{\gamma\sigma}{r_{ij} - a\sigma}\right) \exp\left(\frac{\gamma\sigma}{r_{ik} - a\sigma}\right), \end{aligned} \tag{5}$$

where λ and ε are the interaction potential strength and unit, σ is the length unit, and θ_0 is the reference angle. a represents the three-body cutoff and γ is a damping factor that controls how the potential goes to zero at the cutoff a. The SW potentials has been applied to describe three-body inter-

actions in a number of condensed matter systems in the literature[26–32] and implemented in the LAMMPS MD package.[45] The current implementation of the three-body MS-CG algorithm is to force match the angular dependent part $\lambda_{jik}\varepsilon_{jik}(\cos\theta_{jik}-\cos\theta_0)^2$ and use predetermined parameters for the distance dependent terms. It must be pointed out that the MS-CG procedure adopted to fit the angular term makes no assumptions about its functional form, but it is completely general. However, commonly available MD packages provide only fixed forms for the three-body potential. For this reason, the procedure proposed in this paper will focus on how to obtain the required parameters from a more general fit involving spline functions for the angular part. Presently, the value of the cutoff a is defined by a trial and error procedure and is usually chosen to include the closest neighbor sites. The distance scale σ is chosen to be unity for simplicity. The effect of γ is to tune how the potential goes to zero as the cutoff distance is approached from below. In the present work, a value of 1.2 is chosen for γ. Other γ values were also tested and almost identical results were obtained based on γ varying from 0.6 to 2.4.

The angular dependent term will be expressed as a linear combination of spline functions for the variable θ_{jik} to perform the MS-CG calculation. In order to calculate the derivatives in Eq. (3), an analytical expression for the derivatives must be obtained for certain basis function. Such derivatives need to be continuous. This can be easily achieved using B-splines[46] because the derivatives of B-spline functions are analytically defined.[47]

The procedure for fitting involves two steps. In the first step, a MS-CG calculation is performed using the following basis set:

$$U=\sum_i\sum_{j>i}f^{(2b)}(r_{ij})+\sum_i\sum_{j\neq i}\sum_{k>j}\lambda_{jik}f^{(3b)}(\theta_{jik})\times\exp\left(\frac{\gamma_{ij}\sigma_{ij}}{r_{ij}-a_{ij}\sigma_{ij}}\right)\exp\left(\frac{\gamma_{ik}\sigma_{ik}}{r_{ik}-a_{ik}\sigma_{ik}}\right),\tag{6}$$

where $f^{(2b)}(r_{ij})$ is the two-body part of the potential and $f^{(3b)}(\theta_{jik})$ is the angular part of the three-body potential. For the first step of the fitting procedure, a spline functional form has been chosen for both $f^{(2b)}(r_{ij})$ and $f^{(3b)}(\theta_{jik})$. Splines were used to represent $f^{(3b)}(\theta_{jik})$ in order to estimate the equilibrium position θ_0 in Eq. (5). At this point, it is important to note that there is no reason not to directly use the spline functional form for $f^{(3b)}(\theta_{jik})$ just obtained, except for practical reasons. In the authors' knowledge, no currently available code can manage spline functions to compute the three-body part of nonbonded interactions. For these reasons, a second step is require to correctly fit the available basis set.

A direct fitting to the spline functional form of $f^{(3b)}\times(\theta_{jik})$ with an analytic function, such as $(\cos\theta_{jik}-\cos\theta_0)^2$ introduces non-negligible errors. It is found that the two-body potentials are very sensitive to the corresponding three-body potentials in CG simulations. This means that even for a slight change in the three-body parameters, the two-body parameters need to be modified substantially. This is particularly problematic for the case of using the SW potential in Eqs. (4) and (5). For this reason, a second fitting of the global basis set is required, namely a new MS-CG calculation is performed using the following basis set:

$$U=\sum_i\sum_{j>i}f_{sw}^{(2b)}(r_{ij})+\sum_i\sum_{j\neq i}\sum_{k>j}\lambda_{jik}\varepsilon_{jik}(\cos\theta_{jik}-\cos\theta_0)^2\times\exp\left(\frac{\gamma_{ij}\sigma_{ij}}{r_{ij}-a_{ij}\sigma_{ij}}\right)\exp\left(\frac{\gamma_{ik}\sigma_{ik}}{r_{ik}-a_{ik}\sigma_{ik}}\right),\tag{7}$$

where the SW potential is explicitly used and $f_{sw}^{(2b)}(r_{ij})$ is a spline functional form that in general differs from the $f^{(2b)}\times(r_{ij})$ employed in Eq. (6). The first step is required in order to evaluate the equilibrium angle θ_0, that in kept fixed during the second step of the MS-CG calculation, Eq. (7).

For the water system, the spline result for $f^{(3b)}(\theta_{jik})$ gives a value for θ_0 close to 109°, which is consistent with the tetrahedral structure of water. As a matter of fact, the agreement is very good for TIP3P (Ref. 48) water, whereas for the SPC/E model the best value is around 116°.

The procedure just outlined can be extended to fit more general functional forms such as

$$U(\theta_{jik},r_{ij},r_{ik})=\lambda_{jik}A(\theta_{jik})\exp\left(\frac{\gamma_{ij}\sigma_{ij}}{r_{ij}-a_{ij}\sigma_{ij}}\right)\exp\left(\frac{\gamma_{ik}\sigma_{ik}}{r_{ik}-a_{ik}\sigma_{ik}}\right).\tag{8}$$

Different choices for the distance dependant part of the potential can also be used, making the method here presented useful for the fitting of a broad range of functional forms. In particular, a direct use of the initial $f^{(3b)}(\theta_{jik})$ obtained using a spline functional form in Eq. (6) would require only the first step.

D. Ensemble averages

Ensemble averages are computed using the formula[49]

$$\langle A\rangle_{\mathrm{AA}}=\frac{\int d\vec{p}^{\,n}d\vec{r}^{\,n}A(\vec{r}^{\,n})e^{-\beta U^{\mathrm{AA}}(\vec{r}^{\,n},\vec{p}^{\,n})}}{\int d\vec{p}^{\,n}d\vec{r}^{\,n}e^{-\beta U^{\mathrm{AA}}(\vec{r}^{\,n},\vec{p}^{\,n})}},\tag{9}$$

where $\vec{p}$ are the momenta, $\vec{r}$ are the coordinates, n is the number of particles in the system, and U^{AA} is the all-atom force field. An analogous formula can be written in the case of the CG system

$$\langle A\rangle_{\mathrm{CG}}=\frac{\int d\vec{P}^{N}d\vec{R}^{N}A(\vec{R}^{N})e^{-\beta U^{\mathrm{CG}}(\vec{R}^{N},\vec{P}^{N})}}{\int d\vec{P}^{N}d\vec{R}^{N}e^{-\beta U^{\mathrm{CG}}(\vec{R}^{N},\vec{P}^{N})}},\tag{10}$$

where $\vec{P}$, $\vec{R}$, and N have the similar meaning as in the all-atom case, but refer to the CG system, and U^{CG} is the CG potential. In general, Eqs. (9) and (10) give different values even if computed for the same system. In fact, when a system is CG, some degrees of freedom are removed from the atomistic system and averaged out. This can be understood using as an example case of water studied here. Atomistically, the water molecule has two bonds and one angle. If the water molecule is CG to a one-site particle, then the bonds and the angle are no longer present. This has an effect on the value of A in the CG representation (e.g., the virial part of the pressure and energy). It must be pointed out that corrections to Eq. (10) that take into account the contribution of the removed degrees of freedom can in principle be computed.

An example of this is given in the companion paper in the case of pressure.[50]

Thus, once the atomistic trajectory is known, the contribution of the missing degrees of freedom can be evaluated. However, practically this procedure is not straightforward.[51,52]

For this reason, we adopt a "reverse" approach to the problem. If the CG simulation and the all-atom simulation are correctly sampling the same ensemble, both of them will produce the same ensemble distribution of the CG coordinates $\{\vec{R}^N\}$. In this way, it is possible to use the ensemble of coordinate produced from an all-atom simulation and use it in Eq. (10), where the CG Hamiltonian is used. Since we are only concerned with quantities that are functions of the coordinate positions, we concentrate on the averages generated from the configurational distribution function only.

For this reason, a mapping operator M between the all-atom and CG representations is defined as[21,22]

$$\vec{R} = M(\vec{r}), \tag{11}$$

where $\vec{r}$ are from the atomistic simulation. In this way, the quantity $\langle A \rangle$ can be computed using Eq. (10) even for atomistic data. Thus, a way ensemble averages can be compared between all-atom and CG systems is through Eq. (10). For CG simulations, Eq. (10) is used directly, whereas for atomistic simulations the mapping operator Eq. (11) is applied first. Note that in this way, for a given quantity $A(\vec{R}^N)$ and a given atomistic ensemble $\{\vec{R}^N\}$, its average value $\langle A \rangle$ depends on the specific choice of the CG Hamiltonian $H(\vec{R}^N, \vec{P}^N)$. A *direct comparison with the atomistic average computed using* Eq. (9) *will usually lead to a different result*, as explained previously.

E. Atomistic simulations

As a test case, a cubic box of 1000 rigid SPC/E (Ref. 39) water molecules was chosen. The box edge is 31 Å long. The temperature T was kept constant at 298 K employing a Nose–Hoover thermostat[53–55] with a time constant of 0.5 ps. The equations of motion were integrated according to the leap-frog[10–12] algorithm, with a time step of 2 fs. Electrostatic interactions were computed using the particle mesh Ewald (PME) (Ref. 56) method, with a real-space cutoff of 1 nm. The same cutoff was also employed for the van der Waals interactions. Constraints were employed through the SETTLE algorithm.[57] Configurations were sampled every 1 ps. A simulation of length 2 ns was performed using the MD package GROMACS.[58,59] For calculating the $P\rho$ diagram, different volumes were used. Five different box sizes were employed with edges 31.01, 31.02, 31.03, 31.04, and 31.05 Å long corresponding to a drop in pressure of ~5 bar.

F. CG representation

The CG scheme that will be adopted in the present work consists in mapping the atomistic three-site SPC/E water model into a single CG interaction site. The CG water model does not make any use of electrostatic forces, but employs only short range forces. For comparison, two possible force fields will be presented. In one of them, the CG sites interact only through nonbonded two-body potentials. The second model will employ both a two-body and a three-body potential as will be explained later in this section.

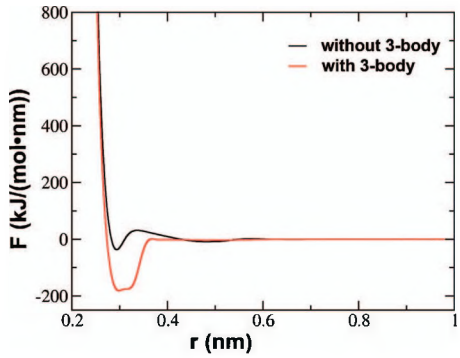

FIG. 1. Interaction forces between CG water sites with (red) and without (black) three-body potential. Only the results for the COM representation are shown.

An important feature of the MS-CG method is that a linear mapping operator is used to transform from the all-atom to CG resolution.[21,22] In the present work, two possible mapping schemes will be employed that map the atomistic water to a single interacting site. In one of them, each water molecule is mapped to a single CG site located at its center of mass (COM representation). Another mapping used in the following calculation is the center of geometry (COG representation). The force field employs tabulated potentials for two-body interactions and a SW potential for the three-body part of the potential.

The two-body part of the interaction forces for the COM representation is shown in Fig. 1, and the MS-CG result with only two-body basis set is shown in the same plot for comparison. It is seen in Fig. 1 that contributions from the two-body forces change dramatically on adding the three-body basis sets to the total potential. This means that the two- and three-body contributions to the total interaction have a strong interdependence, and CG approaches targeted at individual, i.e., either solely two- or three-body distribution functions, are in principle not suitable to parameterize the CG potential in this case. However, the MS-CG approach applies a global optimization strategy and thus is able to deal with this issue by fitting to the total interaction rather than individual two- or three-body contributions. It is also seen from Fig. 1 that the two-body force curve decays to zero at a shorter distance compared to that from the MS-CG results without three-body potential. Because of the fast decay, a short two-body cutoff may be possible for the CG system. The cutoff used for the two-body potential is 6 Å.

For the three-body potential, a cutoff of $a_{ij}=3.7$ Å is used for the COM representation, whereas $a_{ij}=3.9$ Å was found more suitable for the COG model; while $\sigma_{ij}=\varepsilon_{jik}=1$ and $\gamma_{ij}=1.2$ for both models. For the COM representation, $\cos\theta_0=-0.44$ and $\lambda_{jik}=13.135140$; whereas for the COG representation $\cos\theta_0=-0.43$ and $\lambda_{jik}=6.132591$. The parameters were evaluated following the discussion in the previous section. All of the CG simulations were performed using the MD package LAMMPS.[45]

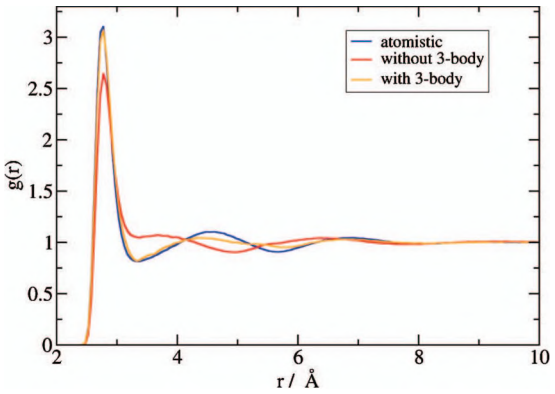

FIG. 2. RDF for the one-site CG model for SPC/E water centered in the center of mass of the interacting particle. The use of an explicit three-body CG potential is seen to improve the first shell of solvation greatly.

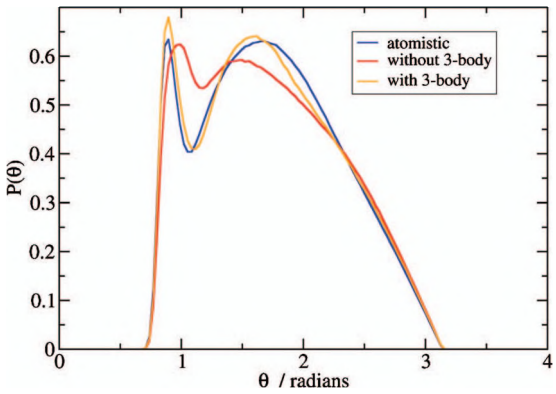

FIG. 4. The ADF for SPC/E water computed inside the cutoff of the three-body CG potential for the case with the center of mass CG representation. The three-body potential is seen to closely reproduce the atomistic distribution.

III. RESULTS AND DISCUSSIONS

A. RDF

The RDF in CG coordinates $g(R)$ is defined as[10]

$$g(R) = \frac{V}{N^2}\left\langle \sum_i \sum_{j\neq i} \delta(\vec{R} - \vec{R}_{ij}) \right\rangle, \tag{12}$$

where V is the volume and R_{ij} is the distance between beads i and j. This function is at the heart of the theory of simple fluids.[42] For this reason, its accurate representation by the CG model should be regarded to be of importance. It can also be shown[60–62] that if pairwise correlations of the system suffice to capture its essential physics, then there is a unique potential (to within a constant) that can reproduce it. However, if higher order correlations are relevant in calculating the properties of the system of interest, a simple two-body approximation for the potential may not capture the essential physics of the system.[51] In Figs. 2 and 3 a comparison between different CG schemes is given. In this comparison, the atomistic RDF is reported in CG resolution, i.e., it represents either the distribution of the COM of the water molecules, or the COG. For each CG scheme, two cases are compared, where only pairwise (i.e., two-body) CG interactions are taken into account during the parameterization, or where both two-body and three-body CG interactions are accounted for. As can be readily seen, the introduction of three-body interactions clearly improves the behavior of the RDF for both CG representations. The first peak of the RDF is now very accurately reproduced, and also the long range structure is reproduced very closely. There is a slight all-atom/CG difference around the second peak, which may be due to the approximations introduced into the three-body CG potential, as described in the methodology section. It should be noted that the three-body potential acts only at very short range, but it is essential for capturing the accurate local structure.[31]

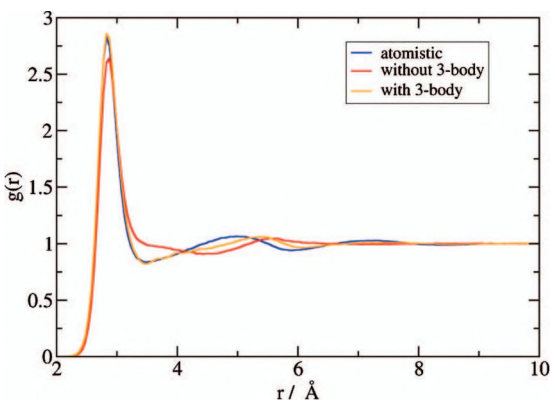

FIG. 3. RDF for the one-site CG model for SPC/E water centered in the center of geometry of the interacting particle.

B. Angular distribution function

As mentioned in the previous section, the RDF is useful to test the quality of the representation of the effective two-body interactions between CG particles. However, due to the fact that three-body interactions are also explicitly taken into account in this work, an appropriate three-body correlation function should be compared as well.[61] To do so, it is defined as

$$P(\theta) = \frac{1}{W}\left\langle \sum_i \sum_{j\neq i} \sum_{k>j} \delta(\theta - \theta_{jik}) \right\rangle, \tag{13}$$

where θ is the angle between a triplet of CG sites. Here, $P(\theta)$ denotes the probability distribution of having the angle θ, and W is a normalization factor. The sum is over all the triplet i, j, k, such that the distances between the jth or kth particle and the particle i are less than 3.7 Å (i.e., the cutoff of the three-body CG potential) or 3.9 Å, for the COM and COG representations, respectively. Roughly, this means that the computation is performed inside the first shell around CG particle i. It is clear from Figs. 4 and 5 that the explicit use of a three-body potential is able to capture more accurately the exact local three-body structure of SPC/E water. A comparison with the result from the simpler pairwise CG approximation reveals that the three-body CG potential creates a better angular distribution around the central CG particle, improving the packing at short range of the interacting particles. Indeed, a two-body CG potential cannot capture the typical angular anisotropy of water, so that the local structure of water appears to be too uniform. The packing properties of the all-atom system are therefore poorly reproduced in the two-body CG potential case.

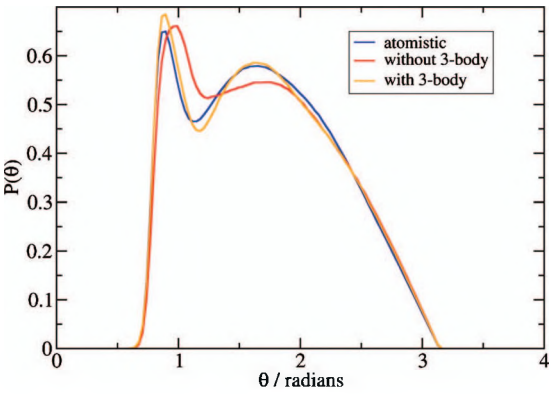

FIG. 5. The ADF for SPC/E water computed inside the cutoff of the three-body CG potential when the center of geometry CG representation is used.

As an approximation of Eqs. (1) and (2) implies that the three-body interaction can be separated into distance and angular dependent terms, through the SW potential functional form used in this work. This separation can eventually lead to a loss of cross correlation between distance and angular dependent contributions. For this reason, in order to validate the approximation in Eq. (2), the cross correlations between distance and angular distributions are calculated in addition to the angular distribution function (ADF), and the results for all-atom and CG configurations are compared in Fig. 6. From Fig. 6, the atomistic result of the two-dimensional probability distribution is reproduced by the CG configurational ensemble, which justifies the usage of the SW potential. Note that in the CG water case the interaction types for r_{ij} and r_{ik} are identical and the probability distributions are plotted as a function of one angle and one distance. There are two red color regions for large probabilities in each of the all-atom/CG plots. These two regions correspond to the two peaks around 50° and 100° in the ADF plots in Fig. 4.

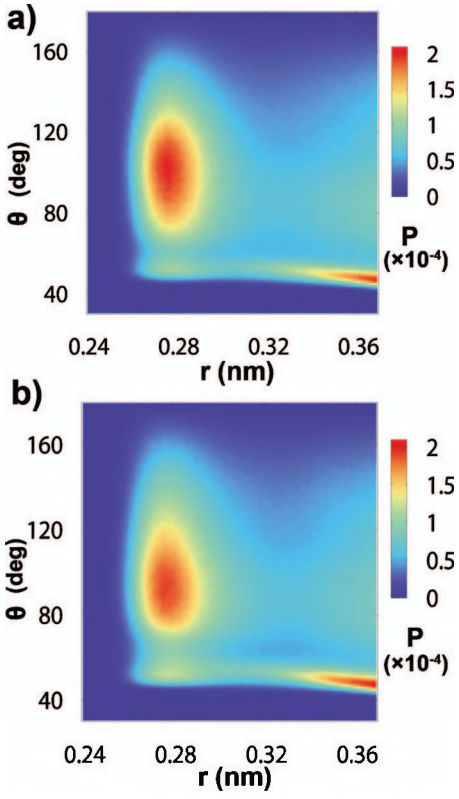

FIG. 6. Probability distributions as a function of two-body distance and three-body angle for each $\{i,j,k\}$ triplet, calculated from: (a) All-atom configurations; (b) CG configurations from CG simulation with both two and three-body CG potentials. The angle θ is θ_{ijk} and the distance r is one of the r_{ij}/r_{ik}. The probability P is calculated for the area $\Delta r \Delta\theta$, with $\Delta r = 0.001$ nm and $\Delta\theta = 1.0°$. Only the results for the COM representation are shown.

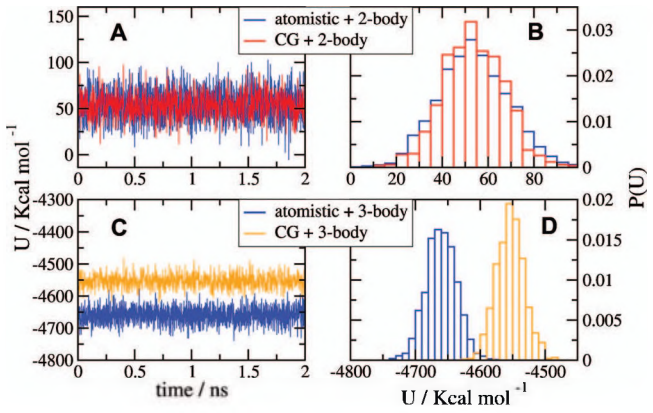

FIG. 7. A comparison of the internal energy between the atomistic and CG systems for SPC/E water. Details of how the comparison was performed can be found in Sec. II. The center of mass case is shown. An analogous plot can be obtained from the case where the geometric center is employed instead. (a) the internal energy is computed using only a two-body CG force field. The red line is an actual CG simulation. The blue line reports the value of the potential energy obtained when the same two-body CG potential is applied to a previous all-atom trajectory in CG resolution. (b) Same as (a), but using a three-body CG potential. (c) The distribution of the data presented in (a). (d) The distribution of the data shown in (b).

C. Ensemble averages

Ensemble averages have been computed according to the procedure described in Sec. II D. The basic concept is that if the CG and the all-atom systems in CG resolution are exploring the same configurational space, a CG thermodynamic function computed using either the CG simulation or the atomistic one should give very similar relative values. In practice, this means that running a CG simulation or an all-atom one would be very similar in the sense explained above. The quantities that are reported here are the potential energy U, the virial part of the pressure $P_{\text{virial}} = [1/(3V)]\Sigma_i^N \Sigma_{j>i}^N F_{ij} \cdot r_{ij}$,[10] and the $P\rho$ diagram, where $\rho = N/V$. All the figures reported refer to the COM case, as the COG case gives analogous results. From Figs. 7 and 8, it is clear that the pressure is better reproduced when using a three-body potential, but the energy is captured in a better way by a simpler two-body approximation. From a similar analysis of TIP3P (Ref. 48) water at the same conditions (data not shown), the trend for pressure is confirmed, but the energy is reproduced with the same precision whether using or not using an explicit three-body potential. Thus, an explicit three-body CG potential seems to improve the pressure systematically, but no clear conclusion can be drawn for the CG potential energy of the system.

The $P\rho$ diagram (Figs. 9 and 10) for CG water presents additional interesting features. In all cases studied, there is an almost constant shift between the fully all-atom and the CG

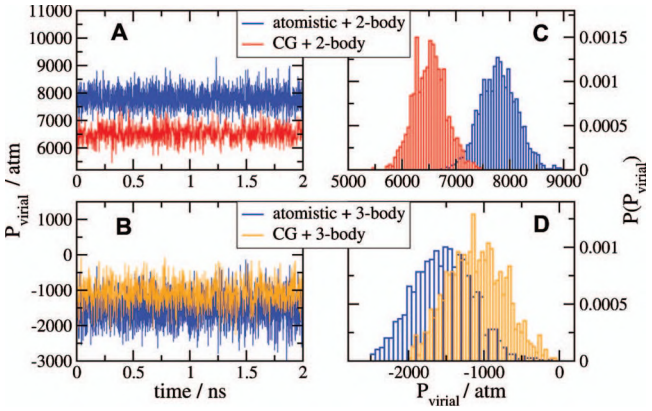

FIG. 8. Comparison of the virial component of the pressure between atomistic and CG simulations for SPC/E water. The case of COM is shown. The case with the center of geometric is analogous. (a) The virial is computed using only a two-body CG force field. The red line is an actual CG simulation. The blue line reports the value of the virial obtained when the same two-body CG potential is applied to a previous all-atom trajectory in CG resolution. (b) Same as (a), but using a three-body CG potential. (c) The distribution of the data presented in (a). (d) The distribution of the data shown in (b).

models. This is confirmed also by the study of TIP3P water model (data not shown). It is obvious from this analysis that the isothermal compressibility is also well reproduced. In fact, this quantity is proportional to the derivative of the $P\rho$ diagram, as it can be written as

$$\kappa_T = \frac{1}{V}\left(\frac{\partial V}{\partial P}\right)_T = \frac{1}{\rho}\left(\frac{\partial \rho}{\partial P}\right)_T = \left(\frac{\partial \ln(\rho)}{\partial P}\right)_T, \tag{14}$$

where V denotes the volume, P is the pressure, and ρ is the density. All of the simulations were performed at the same temperature $T=298$ K. However, CG force fields of the type employed here are temperature dependent. In order to account for this temperature dependence, previous theoretical and computational studies have been performed by Krishna *et al.*[63] to examine and enable the transferability of such CG force fields across different temperatures.

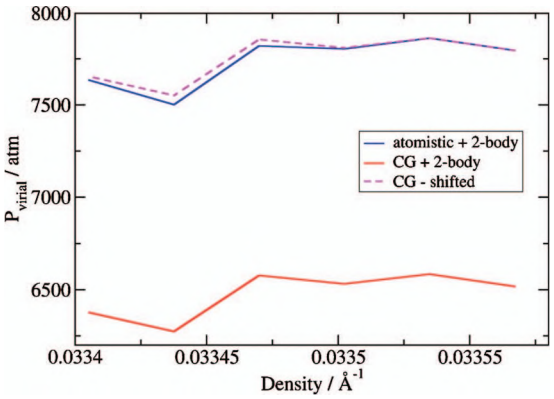

FIG. 9. The virial part of the pressure computed for different densities. The comparison is performed employing the two-body CG potential. For a detailed discussion on how the comparison is performed, see Sec. II D. The curves have the same shape as can be evidenced from shifting the CG simulation data upward (dashed line). The difference is only due to an almost constant shift. Only the center of mass case is shown as the geometric center produces a similar plot.

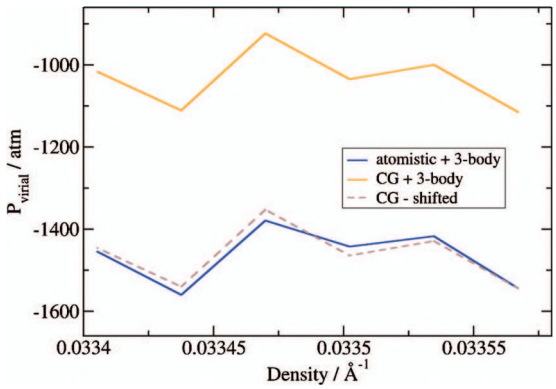

FIG. 10. Same plot as in Fig. 9 but for systems employing an explicit three-body potential. Now the agreement is closer, even though a constant shift can be still observed.

Another important feature of these studies comes from a constant pressure analysis. After performing a series of all-atom constant *NVT* simulations at the same pressure, but at different temperatures, each of these simulations was CG. When the corresponding *NVT* CG simulations were performed, the average pressure was not found to be the same. This finding suggests that the thermal expansion coefficient is different between the atomistic and CG representation of the system. However, considering the possibility of limited statistics, no final conclusion can be drawn about this behavior, which is likely of importance for developing transferable CG force fields. It must be pointed out that describing the correct pressure is a major challenge when building CG potentials. The interested reader can refer to Johnson *et al.*[51] and Louis[52] and references therein. A detailed solution of this problem is found in the companion paper.[50]

D. Computational efficiency

An important feature of the approach proposed here is that simulations with two- and three-body CG potentials in the MD code LAMMPS is significantly faster than all-atom water MD simulations performed with one of the fastest codes presently available for, as can be seen in Table I. The fully atomistic trajectory from LAMMPS was computed using the same parameters as GROMACS, except for the use of Particle-Particle Particle-Mesh (PPPM) (Ref. 64) instead of PME (Ref. 56) for the electrostatics, and SHAKE[65] instead of SETTLE[57] for the rigid molecular constraints. These choices were determined by the fact that the current version of LAMMPS does not support the same algorithms as those in GROMACS. All the simulations were performed on an Intel® Core™ i7 central processing unit at 2.93 MHz with 6 GB of random access memory.

Even though GROMACS is almost five times faster than LAMMPS for fully atomistic MD simulations, a CG simulation with LAMMPS employing a three-body CG potential is still 35% faster than a fully atomistic simulation with GROMACS. It should also be pointed out that in Table I the time step used in all the MD simulations is 2 fs. A larger time step for conventional force field simulations usually causes poor energy conservation,[66] whereas the CG simulations were seen to be up to a stable time step of 8 fs. To evaluate the

TABLE I. Computational efficiency. t_s is the CPU time required to finish the simulation. In the last two columns, this value is normalized with respect to GROMACS' or LAMMPS' t_s. For the three-body case the cutoff employed is given in parentheses.

Kind of simulation and code used	t_s (min)	t_s/t_s (GROMACS)	t_s/t_s (LAMMPS)
Fully atomistic, GROMACS	209	1	0.22
Fully atomistic, LAMMPS	953	4.56	1
CG, no three-body, LAMMPS	15	0.07	0.016
CG, with three-body (3.7 Å), LAMMPS	113	0.54	0.12
CG, with three -body (3.9 Å), LAMMPS	137	0.65	0.14

maximum time step allowed by the CG simulations, multiple simulations were performed with the same initial configuration. Each of them was 10^6 steps long and the different time steps used were 4, 6, 8, and 10 fs. The criteria for this evaluation were the correct reproduction of RDF and ADF, and the overall energy conservation. Usually the structural properties were correctly reproduced, but for the longer time steps the energy conservation became unstable with unphysical spikes appearing on the energy plots for the simulations. In some cases, the effect of such fluctuations caused the simulation to crash. No particular difference in stability was noted when adding an explicit three-body term, leading to the conclusion that the instability is probably due to the two-body part and a time step that was too large.

Such stability of the CG simulation with a longer time step can be easily understood. Atomistic simulations are characterized by short range high frequency motions (which provides a motivation for multiple time step algorithms[66]), whereas in CG models the energy surfaces are smoother in comparison and the CG sites have a higher mass, allowing for greater stability with longer MD time steps. It can therefore be safely concluded that the CG simulation in LAMMPS with the two- and three-body CG terms is effectively a factor of six times faster than the highly optimized all-atom MD code GROMACS for rigid SPC/E water.

IV. CONCLUSIONS

This paper demonstrates that accounting explicitly for three-body contributions to the CG potential can dramatically improve the reproduction of structural properties of the original atomistic system. However, parameterizing multibody CG interactions is a difficult challenge. This paper therefore presents a practical and systematic method for doing this within the MS-CG framework using some additional reasonable approximations. As was shown in Sec. III, the corresponding (algorithmically unoptimized) CG simulation is still significantly more computationally efficient than the currently most efficient fully all-atom MD simulation.

In this work, the importance of having an explicit three-body CG potential has also been clearly affirmed for a simple molecule, such as water.[31] However, the applicability of the present method is broad. It can be used to construct three-body CG force fields for other complex systems such as proteins and nanoparticles. For such CG force fields their major importance will likely be in predicting structural properties.[35,36,38] Three-body potentials are furthermore a critically important component of atomistic force fields for covalently bonded system and noble gases.[26–34]

There are two aspects of the present work which will require further investigation. The first one is the problem of representability.[51] For example, in the MS-CG method it is in principle possible to obtain the best fit to the total CG forces of the system (including now three-body interactions),[20–22] whereas RMC relies on the correct reproduction of RDFs and hence provides a two-body CG potential that in most cases does not capture higher order correlations.[16,17] Unless the system of interest can be well approximated by a two-body CG potential, the two methods will give different answers. A direct comparison of the MS-CG approach that includes three-body CG potentials with other methods such as RMC that focus on the optimal two-body CG potential to reproduce two-body correlations will be a topic of future research.

Another aspect worth further investigation is the role of electrostatic interactions. As pointed out in the work of Molinero and Moore,[31] employing three-body interactions helps to reproduce structural properties, but for many problems the long range electrostatic interactions can play an important role. In water, an explicit representation of its long ranged dipolar nature might be very important, e.g., at interfaces or in other heterogeneous systems, so this feature of the coarse-graining problem will also be a topic of future research.

ACKNOWLEDGMENTS

This research was supported by the National Science Foundation Collaborative Research in Chemistry (Project No. CHE-0628257). Computer resources were provided by the National Science Foundation through TeraGrid computing resources administered by the Pittsburgh Supercomputing Center, the San Diego Supercomputer Center, the National Center for Supercomputing Applications, the Texas Advanced Computing Center, and Argonne National Laboratories. We thank Hans Andersen and Avisek Das for their insightful comments and suggestions regarding this work. The authors also thank Gary Ayton, Valeria Molinero, Vinod Krishna, Ron Hills, Chris Knight, and Mark Maupin for many valuable discussions. L.L. and L.L. contributed equally to this work.

[1] *Coarse-Graining of Condensed Phase and Biomolecular Systems*, edited by G. A. Voth (CRC, Boca Raton, 2008).
[2] *Multiscale Modeling of Soft Matter*, edited by P. Earis; Faraday Discussions (Royal Society of Chemistry, Cambridge, 2010), Vol. 144.

[3] P. G. de Gennes, *Scaling Concepts in Polymer Physics* (Cornell University Press, Ithaca, 1979).
[4] A. Y. Grosberg and A. R. Khokhlov, *Statistical Physics of Macromolecules* (AIP, New York, 1994).
[5] M. Doi and S. F. Edwards, *The Theory of Polymer Dynamics* (Oxford University Press, New York, 1986).
[6] P. G. de Gennes and J. Prost, *The Physics of Liquid Crystals*, 2nd. ed. (Oxford University Press, Oxford, 1995).
[7] V. Tozzini, Curr. Opin. Struct. Biol. **15**, 144 (2005).
[8] C. Wulfing, M. D. Sjaastad, and M. M. Davis, Proc. Natl. Acad. Sci. U.S.A. **95**, 6302 (1998).
[9] A. Arkhipov, P. L. Freddolino, and K. Schulten, Structure **14**, 1767 (2006).
[10] M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids* (Clarendon, Oxford, 1987).
[11] D. Frenkel and B. Smit, *Understanding Molecular Simulation* (Academic, San Diego, 2001).
[12] D. C. Rapaport, *The Art of Molecular Dynamics Simulation*, 2nd ed. (Cambridge University Press, Cambridge, 2004).
[13] W. Tschöp, K. Kremer, J. Batoulis, T. Burger, and O. Hahn, Acta Polym. **49**, 61 (1998).
[14] H. S. Ashbaugh, H. A. Patel, S. K. Kumar, and S. Garde, J. Chem. Phys. **122**, 104908 (2005).
[15] D. Reith, M. Pütz, and F. Müller-Plathe, J. Comput. Chem. **24**, 1624 (2003).
[16] A. P. Lyubartsev and A. Laaksonen, Phys. Rev. E **52**, 3730 (1995).
[17] A. P. Lyubartsev and A. Laaksonen, Lect. Notes Phys. **640**, 2256 (2004).
[18] S. Izvekov and G. A. Voth, J. Phys. Chem. B **109**, 2469 (2005).
[19] S. Izvekov and G. A. Voth, J. Chem. Phys. **123**, 134105 (2005).
[20] W. G. Noid, J. W. Chu, G. S. Ayton, and G. A. Voth, J. Phys. Chem. B **111**, 4116 (2007).
[21] W. G. Noid, J. W. Chu, G. S. Ayton, V. Krishna, S. Izvekov, G. A. Voth, A. Das, and H. C. Andersen, J. Chem. Phys. **128**, 244114 (2008).
[22] W. G. Noid, P. Liu, Y. Wang, J. W. Chu, G. S. Ayton, S. Izvekov, H. C. Andersen, and G. A. Voth, J. Chem. Phys. **128**, 244115 (2008).
[23] F. Ercolessi and J. B. Adams, Europhys. Lett. **26**, 583 (1994).
[24] S. Izvekov, M. Parrinello, C. J. Burnham, and G. A. Voth, J. Chem. Phys. **120**, 10896 (2004).
[25] S. Izvekov and G. A. Voth, J. Phys. Chem. B **109**, 6573 (2005).
[26] M. H. Bhat, V. Molinero, E. Soignard, V. C. Solomon, S. Sastry, J. L. Yarger, and C. A. Angell, Nature (London) **448**, 787 (2007).
[27] A. S. Barnard and S. P. Russo, Mol. Phys. **100**, 1517 (2002).
[28] A. Béré and A. Serra, Philos. Mag. **86**, 2159 (2006).
[29] Z. Q. Wang and D. Stroud, Phys. Rev. B **42**, 5353 (1990).
[30] Z. Q. Wang, D. Stroud, and A. J. Markworth, Phys. Rev. B **40**, 3129 (1989).
[31] V. Molinero and E. B. Moore, J. Phys. Chem. B **113**, 4008 (2009).
[32] F. H. Stillinger and T. A. Weber, Phys. Rev. B **31**, 5262 (1985).
[33] S. Ujevic and S. A. Vitiello, J. Chem. Phys. **119**, 8482 (2003).
[34] J. A. Anta, E. Lomba, and M. Lombardero, Phys. Rev. E **55**, 2707 (1997).
[35] P. Schapotschnikow and T. J. H. Vlugt, J. Chem. Phys. **131**, 124705 (2009).
[36] V. Krishna, G. S. Ayton, and G. A. Voth, Biophys. J. **98**, 18 (2010).
[37] Y. H. Wu, M. Y. Lu, M. Z. Chen, J. L. Li, and J. P. Ma, Protein Sci. **16**, 1449 (2007).
[38] M. R. Ejtehadi, S. P. Avall, and S. S. Plotkin, Proc. Natl. Acad. Sci. U.S.A. **101**, 15088 (2004).
[39] H. J. C. Berendsen, J. R. Grigera, and T. P. Straatsma, J. Phys. Chem. **91**, 6269 (1987).
[40] L. Lu, S. Izvekov, A. Das, H. C. Andersen, and G. A. Voth, J. Chem. Theory Comput. **6**, 954 (2010).
[41] H. S. Ashbaugh and L. R. Pratt, Rev. Mod. Phys. **78**, 159 (2006).
[42] J.-P. Hansen and I. R. McDonald, *Theory of Simple Liquids*, 3rd ed. (Academic, Amsterdam, 2006).
[43] V. Rühle, C. Junghans, A. Lukyanov, K. Kremer, and D. Andrienko, J. Chem. Theory Comput. **5**, 3211 (2009).
[44] W. Smith, T. R. Forester, and I. T. Todorov, *The DL_POLY_2 User Manual* (CCLRC Daresbury Laboratory, Warrington, Cheshire, UK, 2008).
[45] S. J. Plimpton, J. Comput. Phys. **117**, 1 (1995).
[46] L. Lu and G. A. Voth, J. Phys. Chem. B **113**, 1501 (2009).
[47] C. de Boor, *A Practical Guide to Splines* (Springer, New York, 2001).
[48] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, J. Chem. Phys. **79**, 926 (1983).
[49] D. A. McQuarrie, *Statistical Mechanics* (University Science Books, Sausalito, 2000).
[50] A. Das and H. C. Andersen, J. Chem. Phys. 132, 164106 (2010).
[51] M. E. Johnson, T. Head-Gordon, and A. A. Louis, J. Chem. Phys. **126**, 144509 (2007).
[52] A. A. Louis, J. Phys.: Condens. Matter **14**, 9187 (2002).
[53] S. Nosé, Mol. Phys. **52**, 255 (1984).
[54] S. Nosé, J. Chem. Phys. **81**, 511 (1984).
[55] W. G. Hoover, Phys. Rev. A **31**, 1695 (1985).
[56] T. Darden, D. York, and L. Pedersen, J. Chem. Phys. **98**, 10089 (1993).
[57] S. Miyamoto and P. A. Kollman, J. Comput. Chem. **13**, 952 (1992).
[58] E. Lindahl, B. Hess, and D. van der Spoel, J. Mol. Model. **7**, 306 (2001).
[59] D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen, J. Comput. Chem. **26**, 1701 (2005).
[60] R. L. Henderson, Phys. Lett. A **49**, 197 (1974).
[61] J. T. Chayes and L. Chayes, J. Stat. Phys. **36**, 471 (1984).
[62] J. T. Chayes, L. Chayes, and E. H. Lieb, Commun. Math. Phys. **93**, 57 (1984).
[63] V. Krishna, W. G. Noid, and G. A. Voth, J. Chem. Phys. **131**, 024103 (2009).
[64] R. W. Hockney and J. W. Eastwood, *Computer Simulation Using Particles* (McGraw-Hill, New York, 1981).
[65] J. P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen, J. Comput. Phys. **23**, 327 (1977).
[66] R. Zhou, E. Harder, H. Xu, and B. J. Berne, J. Chem. Phys. **115**, 2348 (2001).